Jacobi iterations for canonical dependence analysis

Ronald Phlypo

Jacobi Iterations for Canonical Dependence Analysis

R. Phlypo*

Vision and Brain Signal Processing (ViBS) Research Group, Université de Grenoble / CNRS UMR 5216, Domaine Universitaire, 11, rue des Mathématiques, BP 46, 38402 Saint Martin d'Hères Cedex, FRANCE

Abstract

In this manuscript we will study the advantages of Jacobi iterations to solve the problem of Canonical Dependence Analysis. Canonical Dependence Analysis can be seen as an extension of the Canonical Correlation Analysis where correlation measures are replaced by measures of higher order statistical dependencies. We will show the benefits of choosing an algorithm that exploits the manifold structure on which the optimisation problem can be formulated and contrast our results with the joint blind source separation algorithm that optimises the criterion in its ambient space. A major advantage of the proposed algorithm is the capability of identifying a linear mixture when multiple observation sets are available containing variables that are linearly dependent between the sets, independent within the sets and contaminated with non-Gaussian independent noise. Performance analysis reveals at least linear convergence speed as a function of the number of sweeps.

Keywords: blind source separation, canonical correlation analysis, cumulant tensor, independent component analysis

*Corresponding author
Email address: ronald.phlypo@gipsa-lab.grenoble-inp.fr (R. Phlypo)

Jacobi Iterations for Canonical Dependence Analysis

R. Phlypo*

Vision and Brain Signal Processing (ViBS) Research Group, Université de Grenoble / CNRS UMR 5216, Domaine Universitaire, 11, rue des Mathématiques, BP 46, 38402 Saint Martin d'Hères Cedex, FRANCE

1. Introduction

Many practical measurements result in multiple datasets sharing information through some variables. These variables may be descriptors for some physical quantity and the sets may contain data from different modalities, multiple subjects [1], multiple frequency bins [2, 3], etc. Each of these sets then contains some information that is shared with the other data sets as well as information proper to the respective modality, subject or frequency bin. For instance, one might be interested in simultaneously measuring the positioning of the subject's gaze on a screen and the electro-physiological effect of a change in eye position (e.g., as measured with an electro-oculogram, which are cutaneous electrodes placed in the vicinity of the eyes). Changes in the electrical field measured by the electro-oculogram are not linearly related to changes in gaze direction, but there almost certainly exist a relation. Indeed, some changes in gaze direction can directly be related to an observed change in the electro-physiological measurement. A part from the effects due to a change in gaze direction, the electro-oculogram also records electrical correlates of cerebral activity due to the volume conduction of the brain, skull and skin. The measured gaze direction itself does not, of course, exhibit traces from cerebral activity, although the latter might be contaminated by recording noise due to the system's detection algorithms that are used or its limited spatial accuracy. These errors are obviously not observable on the electro-oculogram. Neither of both may thus act as a pure reference for the other modality. As a consequence this system exhibits all of the above properties, namely that each modality is composed of a set of variables that may explain the information shared between the datasets as well as variables that contain information proper to the recording modality. The latent variables exhibiting the largest dependence between both measurements may thus reveal the electro-physiological activity explained by a change in gaze direction and, vice versa, one may isolate the change of direction that

*Corresponding author
Email address: ronald.phlypo@gipsa-lab.grenoble-inp.fr (R. Phlypo)

is effectively causing a change in the measured electro-physiological potential field.

Our interest lies in the estimation of subspaces of dependent variables from multiple datasets. If we reduce the dependencies between the subspaces to the direct sum of dependencies between the basis vectors spanning these subspaces, we refer to these basis vectors as source components. To estimate those latent variables or source components that explain the information shared between datasets, we may rely on the recently introduced framework of joint blind source separation (JBSS) [4] or Independent Vector Analysis (IVA) [2]. A joint diagonalisation procedure for JBSS has been introduced in [5] and described in more detail in [4]. The goal is to estimate these source components with as little side information as possible, i.e., without disposing of prior information about the distribution family of these components. This is in contrast to the natural gradient procedure introduced in IVA [2], which explicitly assumes the source components to have multivariate Laplace distributions. In this work, we focus on an approach based on non-parametric statistics and will not make any assumption about the source distributions as such closely following the principle of [5, 4]. We will gradually build up our ideas starting from the linear dependencies used in Canonical Correlation Analysis (CCA). Hence, we will refer to the proposed method as Canonical Dependence Analysis (CDA).

We will show how CDA can exploit the structure of the smooth optimisation space, which is actually a manifold. Indeed, suppose we have found a whitening transform for our data that effectively decorrelates the data within each of the datasets, then the manifold over which to optimise is that of the orthogonal matrices [6]. A natural evolution on this manifold outperforms the previously introduced method of [4] based on a second order approximation to the objective function and followed by a refinement step based on re-projected gradients. Empirical results obtained by extensive Monte Carlo simulations, allow us to conjecture what we think might be the two major obstacles encountered in the joint diagonalisation optimisation procedure for JBSS and show how a formulation in the framework of Jacobi iterations and Givens rotations indeed helps to overcome these difficulties. To this end, we will first recapitalise the framework of CCA, of which CDA can be seen as an extension using higher order statistics. We then briefly expose the model of instantaneous linear mixtures in the case of multiple datasets and present the framework of optimisation on the special orthogonal group using Givens rotations and Jacobi iterations [6, 7]. Both he empirical convergence properties of this approach as well as a performance analysis are conducted and a detailed comparison with the state-of-the-art JBSS algorithm [5, 4] allows to reveal both drawbacks and advantages of the JBSS, CCA and CDA methods. We hope this paves the path for future algorithmic developments in this domain.

2. Canonical Correlation Analysis

2.1. Canonical Correlations and Mutual Information

When two datasets are available, Canonical Correlation Analysis (CCA) [8, 9] yields an appropriate solution to finding the information shared between these sets. CCA (also known as the method of *angles between subspaces* [7]) assumes there exists a (non-singular) linear transformation for each of both datasets such that the source components become apparent. The source components, often also called the canonical correlates or the latent variables, are those univariates that share information with a univariate of the other set but with no other variable within the same set. It is straightforward to calculate the shared information for two univariate sets composed of normally distributed components x and y, respectively. For a zero-mean variable $(x, y)^T$ distributed as

$$\mathcal{N}\left((x,y)^T;\mathbf{0},\mathbf{\Sigma}\right) = \frac{1}{2\pi|\det\mathbf{\Sigma}|}e^{-(x,y)\mathbf{\Sigma}^{-1}(x,y)^T}$$

the mutual information between its components is given by

$$\mathrm{MI}(x,y) = -\frac{1}{2}\log\left(1-\rho^2(x,y)\right) \ ,$$

where $\rho^2(x,y)$ is the squared correlation coefficient $E\{x,y\}^2/\left(E\{x^2\}E\{y^2\}\right) = \sigma_{12}^2/(\sigma_{11}\sigma_{22})$. Here, $\mathbf{\Sigma} = (\sigma_{ij})_{i,j}$, $(\cdot)^T$ denotes the transposition and $E\{\cdot\}$ the mathematical expectation operator. The maximum of the mutual information is obtained when the correlation coefficient takes its maximum, i.e., $\rho^2(x,y) \to 1$ implies $\mathrm{MI}(x,y) \to +\infty$. The minimum is obtained at $\rho^2(x,y) = 0$, implying $\mathrm{MI}(x,y) = 0$.

In the multivariate case, we have the additional requirement that no within set dependencies exist. Denote by $\rho\left(x_{i_1}^{[k_1]}, x_{i_2}^{[k_2]}\right), k_1, k_2 \in \{1,2\}$ the correlation coefficient between the i_1-th variable of the set k_1 and the i_2-th variable of the set k_2. The within set dependencies vanish when the components of $\left(x_i^{[k]}\right)_i$ do not share information [10]. As a result, we may write

$$\rho\left(x_{i_1}^{[k_1]}, x_{i_2}^{[k_2]}\right) = 0 \ ,$$

whenever $i_1 \neq i_2$. Imposing that this holds true for whatever values of k_1 and k_2 results in the definition of source components for CCA.

These source components can be obtained by having linear operators acting on the observed random variables $\mathbf{y}^{[k]} = (y_1^{[k]}, y_2^{[k]}, \ldots, y_{N_k}^{[k]})^T$ as

$$\mathbf{x}^{[k]} = \mathbf{Q}^{[k]}\mathbf{y}^{[k]} \ .$$

Only when we limit ourselves to two datasets and no two canonical correlation coefficients are equal, a unique algebraic solution exists to the above problem, which is the solution proposed by Hotelling [9].

Extensions of the above to an arbitrary number of observation sets exist, although one should then relax some of the conditions on the source components [11, 12] [13, and references therein]. In [12], the specific conditions on the source components is found in the requirement of a same ordering of the canonical correlates. In this case, an algebraic solution can be obtained for multiple sets. Unfortunately, the restrictions under which the model works are rarely encountered in practice. When multiple realisations are available under two different conditions, one may also extend the canonical correlation analysis as in [14], although no proof could be found for the offered solution.

All of the aforementioned methods are restricted to the use of (in)dependence measures based on statistics up to order two. As a consequence, they are optimal only when second order statistics are also the sufficient statistics (this is the case for, e.g., normally, log-normally or logistic distributed random variables). In contrast to standard CCA and its extensions discussed above, we prefer to turn to higher order statistics in what follows. Even if second order statistics have proofed their usefulness in many practical applications [8, 9, 12], independence within sets and dependence between sets are measured more efficiently with higher order statistics. Indeed, higher order statistics may reveal dependencies that second order statistics can not detect [15, 2, 5, 4].

2.2. Canonical Correlation Analysis and Least Squares Regression

Linking CCA and Least Squares fitting helps understanding the importance of canonical correlation analysis as a regression technique and the possible applications of CDA and its counterparts JBSS and IVA. Consider two datasets with their associated zero-mean random variables $\mathbf{x}^{[1]}$ and $\mathbf{x}^{[2]}$ with covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, respectively. After whitening of both of the datasets, the optimal linear regression parameters $\mathbf{Q}$ for regression in the least squares sense of $\mathbf{\Sigma}_2^{-1/2}\mathbf{x}^{[2]}$ onto $\mathbf{\Sigma}_1^{-1/2}\mathbf{x}^{[1]}$ are given by $\mathbf{Q} = \arg\min_{\mathbf{H}} E\{\|\mathbf{H}\mathbf{\Sigma}_1^{-1/2}\mathbf{x}^{[1]} - \mathbf{\Sigma}_2^{-1/2}\mathbf{x}^{[2]}\|_2^2\}$. The associated minimum equals $N - \|\mathbf{\Sigma}_1^{-1/2}E\{\mathbf{x}^{[1]}\left(\mathbf{x}^{[2]}\right)^T\}\mathbf{\Sigma}_2^{-1/2}\|_F^2$, where N is the dimension of $\mathbf{x}^{[2]}$. Since the Frobenius norm does not change under orthogonal transformations, we may left and right multiply $\mathbf{\Sigma}_1^{-1/2}E\{\mathbf{x}^{[1]}\left(\mathbf{x}^{[2]}\right)^T\}\mathbf{\Sigma}_2^{-1/2}$ by an arbitrary orthogonal matrix $\mathbf{U}^T$, respectively $\mathbf{V}$. If both these matrices are obtained from its singular value decomposition

$$\mathbf{\Sigma}_1^{-1/2}E\{\mathbf{x}^{[1]}\left(\mathbf{x}^{[2]}\right)^T\}\mathbf{\Sigma}_2^{-1/2} = \mathbf{U}\mathbf{R}\mathbf{V}^T \ ,$$

one finds the canonical correlates $\mathbf{R} = \mathrm{Diag}\left[\left(\rho^2\left(y_i^{[1]}, y_i^{[2]}\right)_i\right)\right]$, where the operator $\mathrm{Diag}(\cdot)$ constructs a matrix with the vector argument on its diagonal. The minimum of the Frobenius norm may now be rewritten as

$$N - \sum_i \rho^2(y_i^{[1]}, y_i^{[2]}) \ ,$$

revealing a minimum regression error when the squared canonical correlations are maximal, i.e., the mutual information is maximal. The goal of this manuscript

is to describe an algorithm which extends this framework for CCA to higher order statistics.

2.3. Source Estimation Using Reference Signals

Some references in literature have attempted to incorporate higher order independence into the regression model, often by reduction of one of the measurement sets to a reference set for the other set. In the extreme case, one of the sets contains a single variable, in which case we refer to this set as a reference variable (signal). A reference signal is called informative if it effectively shares information about the variable(s) of interest with the observation set. The reference signal may appear as an additional, independent observation [16] or as an *a priori* [17]. If the reference signal provides information about the observation noise, this would result in adaptive noise cancellation algorithms. This is probably the most commonly known example and has been treated exhaustively in, e.g., [16]. Main drawbacks of this approach lie in the fact that the noise cancellation focuses on a single reference signal only and restrict the dependency measure to (joint) second order statistics between the data set and the reference signal.

Several contributions in literature provide an extension of the former model to cope with higher order moments to measure the dependencies within the observation set. Lu and Rajapakse [18, 19] have solved the constrained programming problem wherein a single component of the dataset is assumed maximally independent with respect to its complement in the dataset (within dataset independence) [20]. This component is subject to the constraint to lie within the neighbourhood (with respect to some metric) of the reference signal (dependence between sets). Whereas the estimation of the independent component does make use of higher order statistics, the measure that delimits the neighbourhood of the reference signal is restricted to second order cross-statistics in their works. In [21], higher order cross-moments between the reference signal and the output signal are considered, but now the within dataset independence is no longer explicitly exploited. More precisely, the covariance between an even power of the reference signal and an even power of the output signal are considered. Whereas it can be shown that this objective function does yield acceptable results for their application, one should question the restriction to a single cross-statistic of the full (cross-)cumulant tensor, because this is indeed a very approximative measure for dependence. A similar strategy has been followed in [17, 22], where a non-linear function of the data is used in conjunction with a second order optimisation method. It can be shown that this is equivalent to estimating an output signal that shows maximal dependence with a reference signal at higher orders. This method does outperform the second order adaptive filtering method [16] as well as the aforementioned method [21] whenever the samples that have a contribution from the latent variable can be properly identified. Unfortunately, the latter samples can not always be identified easily [22].

The framework of Independent Vector Analysis/Joint Blind Source Separation/Canonical Dependence Analysis, extends the framework of reference signals by attributing a symmetric role to each of the datasets. In addition, the same

measure for within set independence and between set dependence is used. This seems a far more natural setting for joint data analysis than are the combinations of different measures as they have been used in the above references. IVA/JBSS/CDA allow to alleviate the choice of a threshold for the vicinity measure [18, 19], the choice of a single cross-cumulant as a measure for dependence [21] or the estimation of the sample indices for which it is known that the source of interest masks the other sources in amplitude [17].

2.4. Outline of the Manuscript

In what follows, we will first summarise the linear (instantaneous) mixing model. This section also fixes the majority of the notations used in this manuscript and details the constraints under which we will work. The development of this manuscript will be to proceed as for CCA, but with an emphasis on the necessary extensions needed when moving to higher order statistics. This explains why we refer to our approach as Canonical Dependence Analysis, despite the already existing names such as Independent Vector Analysis or Joint Blind Source Separation that may be found in literature. In Section 4, we give a short survey on higher order statistics and cumulants, stressing the advantages of the latter and justifying their presence in CDA. This section also introduces notations for cumulants of $2K-$tuples of variable pairs over different sets. Section 5 will layout the algorithmic approach based on Jacobi iterations and Givens rotations. The results of this algorithmic approach are shown in Section 6, where we compare ourselves to the already existing JBSS algorithm [4] and show the behaviour of our specific algorithm, prooving almost sure convergence based on empirical results. We conclude with the identification of a mixture in the special case of non-Gaussian noise environments. This is an active area of research, and we show that our algorithm obtains good results, without incorporating knowledge other than having two realisations wherein the *sources* of interest are dependent. We conclude the manuscript with a discussion on the results (Section 7), a conclusion and a brief outline of future research directions (Section 8).

3. The linear, instantaneous mixing model and between set interdependencies

In this manuscript, we consider multiple observation sets $\{\mathbf{y}^{[k]}[n], n = 1, 2 \dots N\}, k = 1, 2, \dots, K$ each containing N observations of random variables over the domains $\mathbb{R}^{D_k}$. The random variables $\mathbf{y}^{[k]}$ linearly depend on source variables $\mathbf{s}^{[k]}$ as

$$\mathbf{y}^{[k]} = \mathbf{A}^{[k]}\mathbf{s}^{[k]}, \;\; k = 1, 2, \dots, K \;\; , \tag{1}$$

where we have dropped the index n to simplify notations. We suppose throughout that the matrix representations of the linear operators $\mathbf{A}^{[k]}$ are of full column rank such that their left (Moore-Penrose generalised pseudo-)inverses are

uniquely defined. Denoting these left inverses as $\mathbf{A}^{[k]^\dagger}$, we obviously have the relationship

$$\mathbf{x}^{[k]} = \mathbf{A}^{[k]^\dagger}\mathbf{A}^{[k]}\mathbf{s}^{[k]} = \mathbf{s}^{[k]}, \;\; k = 1, 2$$

where we write $\mathbf{x}^{[k]}$ for the estimate of $\mathbf{s}^{[k]}$. If we furthermore assume that the entries in $\mathbf{s}^{[k]}$ are independent, then so are the entries of $\mathbf{\Pi\Lambda s}^{[k]}$, where $\mathbf{\Pi}$ and $\mathbf{\Lambda}$ represent an arbitrary permutation and a full rank diagonal scaling matrix, respectively.

If the $\mathbf{s}^{[k]}$ are not of the same dimension, we may assume that there exists at least one entry in the $\mathbf{s}^{[k]}$ for which we have that $s_i^{[k_1]}$ depends on $s_j^{[k_2]}$ and this for all (k_1, k_2). Consider the specific case where we may assume that the first I sources show within set dependence over all sets. The source variables can then be ordered such that

$$\exists I, \; 0 < I \leq \min\{D_{k_1}, D_{k_2}\} : \; \forall i \leq I, \; s_i^{[k_1]} \text{ and } s_i^{[k_2]} \text{ are dependent}, \quad (2)$$

whilst $\forall i > I$, $s_i^{[k_1]}$ may be assumed independent with respect to $s_i^{[k_2]}$ and this for all (k_1, k_2). As a consequence, we find that for every couple $(i, j), i \neq j$ we have $s_i^{[k_1]}$ is independent from $s_j^{[k_2]}$, where k_1 and k_2 are not necessarily distinct. This is the exact formulation of CCA (see Section 2), but with dependence replacing the notion of correlation.

Remark that the above formulation is not a necessary condition, but eases the exposition. A more general and relaxed formulation is to assume that for each observation set k, there exists at least one $k' \neq k$ and one index i such that $s_i^{[k]}$ shows dependence with $s_j[k']$ for some j.

In what follows, we show how we may identify $\{\mathbf{s}^{[k]}, k = 1, 2, \ldots, K\}$ as well as $\{\mathbf{A}^{[k]}, k = 1, 2, \ldots, K\}$ in the above model up to the following ambiguities: $\mathbf{x}^{[k]} = \text{blkdiag}(\mathbf{M}_I^{[k]}, \; \mathbf{M}^{[k]}) \; \mathbf{s}^{[k]}, \forall k$. Here, $\mathbf{M}^{[k]}$ represent monomial matrices (also called generalised permutation matrices) of size $(D_k - I) \times (D_k - I)$ and the $\mathbf{M}_I^{[k]}$ are $I \times I$ monomial matrices that share the same permutation factor but may differ in their scaling factor, i.e.,

$$\mathbf{M}_I^{[k]} = \mathbf{\Pi}_I \;\; \mathbf{\Lambda}_I^{[k]} \;\; .$$

The operator 'blkdiag' forms a block diagonal matrix from its arguments. One observes that under the model, the components $s_1^{[k]}, s_2^{[k]}, \ldots s_I^{[k]}$ are bound to be estimated in the same order for all k. In other words, if the estimate $\mathbf{x}_I^{[k_1]} = (x_1^{[k_1]}, x_2^{[k_1]}, \ldots x_I^{[k_1]})^T$ of the set $\mathbf{s}_I^{[k_1]} = (s_1^{[k_1]}, s_2^{[k_1]}, \ldots s_I^{[k_1]})^T$ is a shuffled and rescaled version of $\mathbf{s}_I^{[k_1]}$, then we should encounter the same permutation in the estimate $\mathbf{x}_I^{[k_2]}$ of $\mathbf{s}_I^{[k_2]}$ for all k_2.

If we furthermore jointly represent the source variables as $\mathbf{s} = \left(\mathbf{s}^{[1]^T}, \; \mathbf{s}^{[2]^T}, \; \ldots, \; \mathbf{s}^{[K]^T}\right)^T$, the observations as $\mathbf{y} = \left(\mathbf{y}^{[1]^T}, \; \mathbf{y}^{[2]^T}, \; \ldots, \; \mathbf{y}^{[K]^T}\right)^T$ and the estimates as $\mathbf{x} =$

$\left({\mathbf{x}^{[1]}}^T,\ {\mathbf{x}^{[2]}}^T,\ \ldots,\ {\mathbf{x}^{[K]}}^T\right)^T$, then we have the following relationships:

$$\mathbf{y} = \mathbf{A}\mathbf{s} = \begin{pmatrix} \mathbf{A}^{[1]} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{[2]} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{A}^{[K]} \end{pmatrix} \mathbf{s} \tag{3}$$

and

$$\mathbf{x} = \mathbf{M}\mathbf{s} = \begin{pmatrix} \mathbf{M}_I^{[1]} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^{[1]} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{M}_I^{[2]} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{M}^{[2]} & \ldots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{M}_I^{[K]} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{M}^{[K]} \end{pmatrix} \mathbf{s} \ . \tag{4}$$

An alternative representation can be given by regrouping the dependent source entries by defining an appropriate permutation of the source variables (assuming $D_{k_2} \geq D_{k_1}$ whenever $k_2 > k_1$)

$$\mathbf{s}_\pi = \mathbf{\Pi}_d \mathbf{s} = (s_1^{[1]},\ s_1^{[2]}, \ldots,\ s_1^{[K]},\ s_2^{[1]},\ \ldots,\ s_2^{[K]},\ s_I^{[1]},\ \ldots,\ s_I^{[K]},\ s_{I+1}^{[1]},\ \ldots,\ s_{D_K}^{[K]})^T \ .$$

The independent vector structure then indeed becomes apparent and our estimates now become $\mathbf{x}_\pi = \mathbf{\Pi}_d \mathbf{x} = (\mathbf{\Pi}_d \mathbf{M} \mathbf{\Pi}_d^{-1})\ \mathbf{\Pi}_d \mathbf{s}$. Remark that the permutation $\mathbf{\Pi}_d$ is not uniquely defined, since we may arbitrarily reorder the first KI entries K by K as well as the successive entries $KI+1$ through $\sum_k D_k$, since these operations will not affect the vector independence structure. For some fixed $\mathbf{\Pi}_d$, we observe that the ambiguities under the alternative presentation are given by the monomial matrix $\mathbf{\Pi}_d \mathbf{M} \mathbf{\Pi}_d^{-1}$.

In the remainder of this manuscript, we will assume that the observations $\mathbf{y}^{[k]}$ have been corrected for their mean, decorrelated and normalised (whitened) in their proper observation space, such that their covariance matrices are the identity matrices, in other words $E\{y_i^{[k]} y_j^{[k]}\} = \delta_{ij}$, where the latter is the Dirac delta. This whitening does not impose any restrictions on $E\{y_i^{[k_1]} y_j^{[k_2]}\}$ for $k_1 \neq k_2$ other than $0 \leq |E\{y_i^{[k_1]} y_j^{[k_2]}\}| \leq 1$. Under our working model, we also find that $|E\{s_i^{[k_1]} s_j^{[k_2]}\}| = |\rho(s_i^{[k_1]}, s_j^{[k_2]})|\ \mathbb{I}_{\leq I}(i)$, where $\mathbb{I}_{\leq I}$ is the indicator function[1].

If the observations are not white, whitening operators $\mathbf{W}^{[k]}$ may be defined yielding $\mathbf{y}^{[k]} \leftarrow {\mathbf{W}^{[k]}}^T (\mathbf{y}^{[k]} - \boldsymbol{\mu}_{\mathbf{y}^{[k]}})$, such that $E\{\mathbf{y}^{[k]} {\mathbf{y}^{[k]}}^T\} = \mathbf{I}_{D_k}$. The vector $\boldsymbol{\mu}_{\mathbf{y}^{[k]}}$ contains the mean values of the observations in the k-th dataset. One operator that fulfils this condition is the symmetric matrix square root of the covariance matrix of $\mathbf{y}^{[k]}$. The above whitening step is the decorrelation of the observations in their respective spaces. This is exactly the first step of the CCA.

[1]The indicator function $\mathbb{I}_{\leq I}(i)$ is defined as the function that maps its argument i to 1 if it fulfils the condition $i \leq I$, otherwise i is mapped to 0. This function is also known as the Heaviside step function or the hard thresholding function.

Let us assume from here on that the observations have been whitened. The estimates are then related to the observations as $\mathbf{x}^{[k]} = {\mathbf{Q}^{[k]}}^T \mathbf{y}^{[k]}$, where the $\mathbf{Q}^{[k]}$ belong to the group of (special) orthogonal linear operators [6]. For whitened observations, the $\mathbf{A}^{[k]}$ may thus be assumed to belong to the group of (special) orthogonal linear operators. CCA will find those $\mathbf{Q}^{[k]}$ for which the between set correlations are maximal. Because maximal correlation is not necessarily equivalent to maximal dependence – except for some limited class of distributions – we will here focus on their higher order equivalents, the cross-cumulants.

4. Cumulants for Multiple Observation Sets

Let us abuse the naming convention and refer to $\mathbf{y}^{[k_2]}, k_2 \neq k_1$ as the momentary fixed reference sets, containing information about the sources we would like to estimate from a given set $\mathbf{y}^{[k_1]}$. Remark that the fixed set is not given as a set of independent variables, but as a set of raw observations and thus possibly also contain non-informative contributions with respect to the variables of interest. Let us have a closer look at how these *fixed* observation sets interplay with the observation set $\mathbf{y}^{[k_1]}$.

Thereto, let us first introduce some further notations. General cumulants of order R will be denoted as $\mathrm{Cum}\{y_{i_1},\ y_{i_2}, \ldots,\ y_{i_R}\}$, where indices may be repeated. Adopting the notation by Kendall [23] for cumulants of a $T-$tuple $\mathbf{y} = (y_{i_1},\ y_{i_2},\ y_{i_3}, \ldots,\ y_{i_T})^T$ with no two repeated indices, we write

$$\kappa^{\mathbf{y}}_{p_1 p_2 \ldots p_T} = \mathrm{Cum}\{\underbrace{y_{i_1}, \ldots,\ y_{i_1}}_{\times p_1},\ \underbrace{y_{i_2}, \ldots,\ y_{i_2}}_{\times p_2}, \ldots,\ \underbrace{y_{i_T}, \ldots,\ y_{i_T}}_{\times p_T}\}\ . \qquad (5)$$

Define the $2K$-tuple $\mathbf{y}_{[i,j]} \stackrel{\mathrm{def}}{=} (y_i^{[1]},\ y_j^{[1]},\ y_i^{[2]},\ y_j^{[2]}, \ldots,\ y_i^{[K]},\ y_j^{[K]})$. To adapt the above notation to multiple sets we will define the following notation for the joint cumulants of this $2K$-tuple:

$$\kappa^{\mathbf{y}_{[i,j]}}_{p,q,r,s}(k_1) = \sum_{k \neq k_1} \mathrm{Cum}\{\underbrace{y_i^{[k_1]}, y_i^{[k_1]}, \ldots, y_i^{[k_1]}}_{\times p}, \underbrace{y_j^{[k_1]}, y_j^{[k_1]}, \ldots, y_j^{[k_1]}}_{\times q}, \underbrace{y_i^{[k]}, y_i^{[k]}, \ldots, y_i^{[k]}}_{\times r} \underbrace{y_j^{[k]}, y_j^{[k]}, \ldots, y_j^{[k]}}_{\times s}\}$$

where k_1 is the set that momentarily gets our attention. For our zero-mean and whitened observations $\mathbf{y}^{[k]}$ and fixing $k_1 = 1$, we then have, for all i and j,

$$\begin{aligned}
\kappa^{\mathbf{y}_{[i,j]}}_{2,0,1,0}(1) &= \sum_{k=2}^{K} E\{y_i^{[1]} y_i^{[1]} y_j^{[k]}\} \\
\kappa^{\mathbf{y}_{[i,j]}}_{1,1,1,1}(1) &= \sum_{k=2}^{K} E\{y_i^{[1]}\ y_j^{[1]}\ y_i^{[k]}\ y_j^{[k]}\} - E\{y_i^{[1]}\ y_i^{[k]}\} E\{y_j^{[1]}\ y_j^{[k]}\} \\
&\qquad - E\{y_i^{[1]}\ y_j^{[k]}\} E\{y_j^{[1]}\ y_i^{[k]}\} - E\{y_i^{[1]}\ y_j^{[1]}\} E\{y_i^{[k]}\ y_j^{[k]}\}
\end{aligned}$$

where we have used the fact that our observations have been whitened. For notational simplicity, we will drop the indices (i, j) in the superscript as well as the reference to the set of interest if this does not lead to ambiguity.

Taking any pair of indices (i,j), $i,j \leq \min\{D_k\}$ defines a tuple for which it is easy to verify that under the working model the only non-zero entries in the $R-$th order source cumulant tensor are given by $\kappa^{\mathbf{s}}_{r_1,0,r_2,0}$, $\kappa^{\mathbf{s}}_{0,r_1,0,r_2}$, for any pair $r_1, r_2 \geq 0$ for which $r_1 + r_2 = R$ (and this for any set k_1 of interest). For $\mathbf{x}$ to be an estimate of the sources we thus require that the entries $\kappa^{\mathbf{x}}_{r_1,r_2,0,0}, \kappa^{\mathbf{x}}_{0,0,r_1,r_2}, \kappa^{\mathbf{x}}_{r_1,0,0,r_2}$ and $\kappa^{\mathbf{x}}_{0,r_1,r_2,0}$ vanish for any r_1, r_2 strictly positive and $r_1+r_2 = R$. In addition, we find that $\kappa^{\mathbf{x}}_{r_1,r_2,r_3,r_4} = 0$ for all $r_1, r_2, r_3, r_4 (r_1 + r_2 + r_3 + r_4 = R)$ amongst which at maximum one is zero. In other words, for any tuple $\mathbf{x}_{[i,j]}$, the sum of the squares of all these entries should vanish. Since we also have that the sum of the squares of all entries in the cumulant tensor of a given degree is an invariant under orthogonal transformations of the variables [24, 6], the aforementioned minimisation problem is equivalent to a maximisation of the sum of the squares of the entries $\kappa^{\mathbf{x}}_{r_1,0,r_2,0}$ and $\kappa^{\mathbf{x}}_{0,r_1,0,r_2}$ for which $r_1 + r_2 = R$ over the group of orthogonal transformations acting on $\mathbf{x}^{[k]}$ and this for every reference set k_1 in the calculation of the cumulants for the tuples $\mathbf{x}_{[i,j]}$. It is worth noting that for $k \in \{1,2\}$ one may indeed pass from $\kappa^{\mathbf{x}}_{r_1,0,r_2,0}$ to $\kappa^{\mathbf{x}}_{0,r_2,0,r_1}$ by a simple change of the set of interest.

Let us focus on the case of two observation sets. All of the above holds when $i < j \leq \min\{D_1, D_2, I\}$. But, what if the variables are supposed independent, i.e., the variable $s^{[1]}_j$ does not depend on the variable $s^{[2]}_j$ ($i \leq I < j \leq \min\{D_1, D_2\}$), or, the variable $s^{[2]}_j$ does not have a counterpart ($i \leq D_1 < j \leq D_2$)? In the former case, the model puts restrictions on $\kappa^{\mathbf{x}}_{r_1,0,r_2,0}$ and $\kappa^{\mathbf{x}}_{0,r_1,0,r_2}$, since they now also need to be zero with $r_1 + r_2 = R$ for any $r_1, r_2 > 0$. For the latter case, things are slightly more complicated. To ease the exposition, we choose to formally extend the first set by introducing phantom variables $\square^{[1]}_i$ and $D_1 < i \leq D_2$, completing as such (at least formally) the smallest observation set to be of the same size of the largest dataset. In practice, this may be implemented by adding extra variables with distribution δ_0, although a such naive implementation would result in unnecessary computations. For the entries in the cumulant tensor of order R, we will use the same symbol to denote that we are dealing with a phantom variable rather than an observable, which will lead to the following equivalent notations $\kappa^{(x^{[1]}_i, \square^{[1]}_j, x^{[2]}_i, x^{[2]}_j)}_{r_1,r_2,r_3,r_4} \equiv \kappa^{(x^{[1]}_i, \square^{[1]}_j, x^{[2]}_i, x^{[2]}_j)}_{r_1,\square,r_3,r_4} \equiv \kappa^{\mathbf{x}}_{r_1,\square,r_3,r_4} \equiv \kappa^{\mathbf{x}}_{r_1,r_2,r_3,r_4}$, and $(x^{[1]}_i, \square^{[1]}_j, x^{[2]}_i, x^{[2]}_j) \equiv \mathbf{x}_{[i,j]} \equiv \mathbf{x}$. The last equivalence relations are very implicit representations, but they will allow us to maintain a general notation, without explicit bookkeeping of the phantom variables. Remark that $\square$ has no attributed value and thus the summing restriction in our example reduces to $r_1 + r_3 + r_4 = R$, instead of $r_1 + r_2 + r_3 + r_4 = R$ for the general case.

5. Algorithmic Approach

In this manuscript we show how the use of Givens' rotations and Jacobi iterations can be used successfully in the updating of the estimates of the orthogonal matrices. The rotation matrix $\mathbf{Q} = \widehat{\mathbf{MA}^{-1}}$ can be written as

blkdiag($\mathbf{Q}^{[1]}$, ..., $\mathbf{Q}^{[K]}$), which means that only pairs (i, j) should be considered that belong to the same observation space. Indeed, we actually have an observation space $\mathbb{R}^{D_1} \oplus \mathbb{R}^{D_2} \oplus \ldots \oplus \mathbb{R}^{D_K}$, where the rotations $\mathbf{Q}^{[k]}$ act on the respective subspaces. Now, for an index pair (i, j) with $i < j \leq \max_k(D_k)$ we have the associated quadruple $\mathbf{x}_{[i,j]}$. Thus, once the pair (i, j) and the current set of interest k_1 has been fixed, we may limit our focus to the maximisation of the squares of the entries $\kappa^{\mathbf{x}}_{r_1,0,r_2,0}(k_1)$ and $\kappa^{\mathbf{x}}_{0,r_1,0,r_2}(k_1)$ for $r_1 + r_2 = R$ in the cumulant tensor of order R, as we have derived above.

Each rotation matrix in $\mathrm{SO}(D_k)$ can be expressed as a (non-unique) series of $D_k(D_k-1)/2$ planar rotations as $\mathbf{Q}^{[k]} = \prod_{i,j>i} \mathbf{Q}^{[k]}_{ij}(\theta^{[k]}_{ij})$. As a consequence, we may write the cumulants of the output $\mathbf{x}$ as a function of the cumulants of the observations $\mathbf{y}$ and the parameter set $\{\theta^{[k]}_{ij}\}$ [6]. Since rotations in spaces of dimension larger than two are not commuting, the set $\{\theta^{[k]}_{ij}\}$ is necessarily an ordered set.

To simplify notations, we use the updating scheme as described in Algorithm C.1. The updating scheme calculates the current optimal planar rotation by maximisation of a contrast function depending solely on $\theta^{[k]}_{ij}$, where k refers to the current set of interest. The observations are subsequently updated accordingly and we continue our iterations by taking the next index triple (i, j, k) determining $\mathbf{y}_{[i,j]}$ and $\theta^{[k]}_{ij}$. Because of the non-uniqueness of representation, once all triples have been exhausted, we need to re-initiate a sweep over all triples. This needs to be repeated until convergence.

[Table 1 about here.]

The updates involve trigonometric functions that may be expressed as rational polynomials in $\tan\left(\theta^{[k]}_{ij}\right)$. Setting the derivatives of the sum of squared cumulants to 0, we obtain a polynomial in $\tan\left(\theta^{[k]}_{ij}\right)$ of degree $2R$ that is reducible. If the polynomial has a symmetry, the reduced form of the polynomial is of order R and analytic solutions are available for its roots whenever $R \leq 4$. Unfortunately, if $i \leq I$ the polynomial looses its symmetry and seems irreducible. Thus, analytic solutions are not available. The rooting of these polynomials is done numerically by using the eigenvalue decomposition of the companion matrix (see Appendix B). For many datasets $I \approx \min_k\{D_k\} \ll \max_k\{D_k\}$, resulting in an affordable number of pairs associated with a polynomial of the latter kind (for example, when a single reference variable is available for a set of observations). The coefficients of the polynomials as a function of the observation cumulants for a given quadruple $\mathbf{x}_{[i,j]}$ are given in Appendix A. We observe that quadruples for which $I < i < j$ result in the polynomials underpinning the CoM2 algorithm for Independent Component Analysis [6].

Remark 1. If we follow the above algorithmic construction for statistics of order two, we have $2R = 4$ and thus fourth order polynomials only. This leads to closed form analytic solutions at each iteration. Indeed, the above

reasoning reduces to the typical form of canonical correlation analysis for which an algebraic solution is known to exist.

Canonical Dependence Analysis can thus be summarised in the following algorithm:

[Table 2 about here.]

6. Results on Simulated Datasets

6.1. Performance Evaluation

In order to evaluate the outcome of our simulations we use performance indices based on the Moreau-Amari [25, 26] performance index, defined as

$$\mathrm{PI}(\mathbf{G}) = \frac{1}{2N(N-1)} \left[\sum_{i=1}^{N} \left(\frac{\sum_{j=1}^{N} |g_{ij}|}{\max_d |g_{id}|} - 1 \right) + \sum_{n=1}^{N} \left(\frac{\sum_{m=1}^{N} |g_{mn}|}{\max_f |g_{fn}|} - 1 \right) \right] \quad . \tag{6}$$

To evaluate the fixed permutations for the first I source variables in both observation sets, we introduce the following connection matrix (only valid for $K = 2$): $\mathbf{P} = \left(\mathbf{G}^{[1]}\right)^T \mathbf{G}^{[2]} = \left(\mathbf{A}^{[1]}\right)^T \mathbf{Q}^{[1]} \left(\mathbf{Q}^{[2]}\right)^T \mathbf{A}^{[2]}$, matching the source estimates of the first observation set to those of the second observation set. Our model requires $\mathbf{P}$ to be of the form

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\sigma}_{I \times I} & \mathbf{0}_{I \times D_1} \\ \mathbf{0}_{D_2 \times I} & \boldsymbol{\Pi}_{\mathbf{G}(D_2 - I) \times (D_1 - I)} \end{pmatrix} \quad , \tag{7}$$

where $\boldsymbol{\Pi}_{\mathbf{G}}$ represents an arbitrary permutation matrix and $\boldsymbol{\sigma}$ is a diagonal $I \times I$ matrix with $+1$'s and -1's on its diagonal. We propose the following adapted performance index (for $D_1 = D_2$):

$$\mathrm{PI}_{\mathrm{DCA}}(\mathbf{P}) = \frac{\|\mathrm{zdiag}(\mathbf{P}_{11})\|_1 + \|\mathbf{P}_{12}\|_1 + \|\mathbf{P}_{21}\|_1 + C_1 \ \mathrm{PI}(\mathbf{P}_{22})}{C_2} \quad , \tag{8}$$

where $C_1 = 2(D_1 - I)(D_1 - I - 1)$ and $C_2 = D_1(D_1 - 1)$ are constant scaling factors, $\mathrm{zdiag}(\cdot)$ results in a zeroing of the diagonal elements of its matrix argument and $\| \cdot \|_1$ is the sum of the absolute values of the entries of its argument. This measure proofs essentially useful whenever $I < D_1 = D_2$ and thus a distinction should be made between the dependent subspace and the independent subspace.

Another useful measure, is that of the joint Inter Symbol Interference ($\mathrm{ISI_J}$) index as introduced in [12]:

$$\mathrm{ISI_J}(\{\mathbf{Q}^{[k]}\}) = \mathrm{PI}\left(\sum_k \left| \mathbf{Q}^{[k]} \mathbf{A}^{[k]} \right| \right) \quad , \tag{9}$$

where $\mathrm{PI}(\cdot)$ again refers to the Moreau-Amari performance index PI of Equation (6). This measure proofs essentially useful whenever $D_k = I$ for all k.

Remark 2. When we compare to the JBSS algorithm, we will always refer to the version exploiting all N^2 slices of the cumulant tensor. If not explicitly stated, it is assumed that the quadratic cross cost has been minimised followed by a refinement step based on a gradient descent on the whole objective function. This is the fairest comparison with respect to the CDA method, which acts on all of the entries in the fourth order cumulant tensor.

A summary of the experiments and their results is given in Table C.3.

[Table 3 about here.]

6.2. Multivariate Laplace distributions

A first step in the validation of the proposed algorithm is to repeat part of the experiments proposed in [4]. The experiment of interest is that of the observed instantaneous mixtures of multivariate Laplace distributed variables. Multivariate Laplace variables can be generated starting from independently exponentially distributed $\gamma_i \sim \text{Exp}(1)$ and independently normally distributed $z_i^{[k]} \sim \mathcal{N}(0,1)$. If we pose $s_i^{[k]} = \sqrt{\gamma_i} z_i^{[k]}$, we have a multivariate Laplace distribution for $\mathbf{s}$ [2]. Since the multivariate Laplacian is a symmetric distribution, we have that all odd order cumulants vanish. In addition, we find $E\{\mathbf{s}\mathbf{s}^T\} = \mathbf{I}$, which is a consequence of the independence of all the variables γ_i and $z_i^{[k]}$ involved, from which it results that CCA can not be used here to identify the mixtures or the canonical correlates.

We have the following expression for the fourth order cumulants:

$$\begin{aligned}&\text{Cum}\{s_{i_1}^{[k_1]}, s_{i_2}^{[k_2]}, s_{i_3}^{[k_3]}, s_{i_4}^{[k_4]}\} = \\ &\delta_{i_1,i_2}\delta_{i_3,i_4}\delta_{i_1,i_3}\left(\delta_{k_1,k_2}\delta_{k_3,k_4} + \delta_{k_1,k_3}\delta_{k_2,k_4} + \delta_{k_1,k_4}\delta_{k_2,k_3}\right) \in \{0,1,3\} \ ,\end{aligned}$$

indicating the possible usefulness of fourth orders statistics in an estimation algorithm for the identification of both the mixtures and the latent variables.

6.2.1. Convergence Analysis

Let us first focus on the empirical convergence of the algorithm on this dataset. Thereto we evaluate the outcome of 1000 Monte Carlo runs through the joint ISI performance index. Each Monte Carlo realisation draws 10^4 samples for γ_i and $z_i^{[k]}$ as defined above and a random orthogonal mixing matrix $\mathbf{A}$ structured as in Equation (3). The evolution of the performance index over 10 sweeps is shown in Figure C.1 for $D_1 = D_2 = I = 3$.

[Figure 1 about here.]

To analyse the convergence for varying dimensions $D_1 = D_2 = I$, we submitted the performance indices obtained from consecutive sweeps to the non-parametric statistical test of Kolmogorov-Smirnov (KS). The KS test determines whether two distributions differ significantly (H1) against the null hypothesis (H0) that the performance indices in both consecutive sweeps have the same

underlying distribution. A difference in distribution is judged significant if the p-value – the probability of observing the obtained value for the test statistic under H0 – is sufficiently small[2]. In Table C.4 we show the p-values when comparing the distributions of the performance indices over consecutive sweeps. We observe that for a given number of sources $D_1 = D_2 = I$, there exists some t at which convergence in distribution can be assumed, since no longer can we attribute any significance to the observed difference ($p \approx 1$).

[Table 4 about here.]

6.2.2. Performance Comparison on Theoretical Cumulant Tensor

A first performance study reports on both the JBSS algorithm in its different forms and the Jacobi/Givens-based CDA algorithm when we assume to have the population statistics. In other words, the algorithm will take the theoretical cumulant values as their input. We limit the study to the case of two observation sets, but results may be extrapolated to more sets as in [4]. We exclude the IVA algorithm from the comparison because of the local convergence and permutation issues raised in [4, 27], since these issues indeed result in consistent inferior performance for all sample sizes reported.

In this study, we construct a smooth path from the exact unmixing matrix for the second set to a matrix at distance $\pi/2$ in $\mathrm{SO}(D_2)$ (see Appendix C for its parametrisation and its generation). The initialisation of the demixing matrix of the first set is taken to be the exact unmixing matrix. But, this choice is of little to no importance, since, in both algorithms the demixing matrix of the first set is updated first, based on the current estimate of the demixing matrix of the second set. It is only when the gradient search has been isolated from the cross-cost minimisation initialisation [4] that this initialisation will indeed effect the performance. Because we have chosen to start at the exact unmixing matrix for the first set and since we are moving smoothly away from the exact unmixing matrix for the second set (i.e., no abrupt permutation), this should not have any negative effect on the performance.

In Figure C.2, we show the mean of 100 Monte Carlo runs for 100 homogeneously spread values of the geodesic distance $\varepsilon \in [0, \pi/2]$. Each Monte Carlo Run repeats all values of ε for a given random realisation of the sources $\mathbf{s}$ (as detailed above) and generates a random orthogonal mixing matrix $\mathbf{A}$ as given in the model of Equation (3). The dimensions of the sets are $D_1 = D_2 = I = 3$.

[Figure 2 about here.]

6.2.3. Performance comparison for Finite Large Sample Size

The same experiment has been repeated, but now for a limited number of samples drawn from the distributions specified in Subsection 6.2.2. The number

[2]One refers also to this probability as the probability that the operator makes an error upon rejection of the hypothesis H0.

of samples is 10^4, introducing moderate bias and variance in the estimates of the cumulants up to order four. The results of this experiment are shown in Figure C.3.

[Figure 3 about here.]

6.3. Non-Gaussian noise perturbations

Often in practice we encounter an environment in which the complementary (nuisance) sources or the noise do not have a Gaussian distribution. In this case, simple source separation algorithms are not appropriate to estimate the mixing matrix. When a second observation set is available wherein the variables of interest are present, but now with independent nuisance sources or noise, then we might expect that both JBSS, CCA and CDA may help in recovering the mixing matrix. To evaluate the performance in this setting, we concentrate on the following generative model, extending the model of Equation (1) to:

$$\mathbf{y}^{[k]} = \mathbf{A}^{[k]}\mathbf{s} + \nu\, \mathbf{B}^{[k]}\boldsymbol{\eta}^{[k]} = \mathbf{A}^{[k]} \left[\mathbf{s} + \nu \left(\mathbf{A}^{[k]}\right)^{-1} \mathbf{B}^{[k]}\boldsymbol{\eta}^{[k]} \right] = \mathbf{A}^{[k]}\tilde{\mathbf{s}}^{[k]} \ , \ k = 1, 2 \ , \tag{10}$$

where $\mathbf{s}$ has unit variance non-Gaussian entries, I of which are common to both observation sets. The remaining $D_k - I$ variables are specific to the observation set. ν is a free variable that can be used to obtain a desired signal-to-noise ratio. We choose to limit our case studies to $\nu = 0$ or $\nu = 1$. $\boldsymbol{\eta}^{[k]}$ contains nuisance sources or noise, which may be of Gaussian or non-Gaussian nature. In the simulations, all random variables ($\mathbf{s}^{[k]}$'s and $\boldsymbol{\eta}^{[k]}$'s) are identically and independently distributed with entries drawn either from the uniform or the doubly exponential (Laplacian) distribution. Whether an entry in the random variables is uniform or Laplacian in its marginal distribution is determined by drawing a random variable from a Bernoulli distribution ($p = 0.5$) and attributing either one of the outcomes to one of the distributions. A Gaussian distribution for $\boldsymbol{\eta}^{[k]}$ has not been considered here, since these would not effect the higher order cumulants.

The random variable $\tilde{\mathbf{s}}^{[k]}$ has entries $s_i^{[k]} + \nu \left(\mathbf{c}_i^{[k]}\right)^T \boldsymbol{\eta}^{[k]}$, where $\mathbf{c}_i^{[k]}$ is the $i-$th column of $\nu \left[\left(\mathbf{A}^{[k]}\right)^{-1} \mathbf{B}^{[k]}\right]^T$. This random variable thus contains dependent entries whenever $\nu \neq 0$ and the columns of $\mathbf{A}$ and $\mathbf{B}$ are not aligned. By choosing the $\boldsymbol{\eta}^{(k]}$ non-Gaussian, $\left(\mathbf{c}_i^{[k]}\right)^T \boldsymbol{\eta}^{[k]}$ will be non-Gaussian. In fact, its distribution would tend to that of a normal distributed variable only for large D_k (by the central limit theorem). For reasonably small D_k its distribution is at least as Gaussian as the most Gaussian among the marginal distributions of $\boldsymbol{\eta}^{[k]}$, albeit rarely Gaussian.

6.3.1. Convergence

The empirical convergence for 1000 Monte Carlo runs ($D_1 = D_2 = I$) for 10^4 samples and $\nu = 1$ (nuisance sources and sources of interest with equal power)

are shown in Figure C.4. At each Monte Carlo iteration, all latent variables are randomly drawn as defined above. The orthogonal mixing matrices are drawn randomly following the model of Equation (3).

[Figure 4 about here.]

To test the convergence for varying dimensions $D_1 = D_2 = I$, we submitted the obtained performance indices of consecutive sweeps to a KS-test, see higher for details. The results are reported in Table C.5.

[Table 5 about here.]

6.3.2. Performance Analysis

We here investigate the sensitivity of the algorithm to a mismatch in the prior on I and judge its performance with respect to both the JBSS and CCA method. Indeed, we may suspect that the results undergo an influence of whether I (the number of common sources in the realisations) can sufficiently well be approximated through $\hat{I}$ (the value for I that is used in the algorithm). Results for $I = 1, 2, 3$ and $\hat{I} = 0, 1, 2, 3$ (remark that $\hat{I} = 0$ corresponds to the regular CoM2 ICA algorithm [6]) acquired from 1000 Monte Carlo runs are given in Figure C.5 and are compared to the results obtained by JBSS and CCA. Performance is measured both through the performance index $\mathrm{PI}_{\mathrm{CDA}}(\mathbf{P})$ as given in Equation (8) and the joint ISI of Equation (9).

[Figure 5 about here.]

7. Discussion

In this manuscript, we have introduced an original approach to canonical dependence analysis, based on Jacobi iterations. Whilst the concept of CDA is in its principle not different from that of JBSS, it has been shown that the algorithmic approach taken in this paper has certain advantages over the already existing methods in literature.

7.1. Convergence of the Jacobi algorithm

First of all, we have established the empirical convergence properties. Figure C.4, shows the evolution of the measure $\mathrm{PI}_{\mathrm{CDA}}$ with respect to the iteration number. The results in Tables C.4 and C.5 for different random variable sizes $D_1 = D_2 = I \in \{2, 3, 5, 10\}$ show that the performance indices indeed converge in distribution over the iterations. In addition, the final distribution of the performance measures (for $t \to +\infty$) shows an acceptable performance for our algorithm (Figure C.1).

7.2. Performance on Synthetic Datasets

Indeed, as shown in Figures C.2 and C.3, the CDA approach seems to slightly outperform JBSS when we consider multivariate Laplace distributions. This is the case for both the population statistics as well as a limited sample size of 10^4 samples. As has been reported in [4], the optimisation function may contain a lot of local optima, resulting in the poor performance of the gradient only approach and the limited increase in performance when the gradient step is used as a refinement step following the optimisation of the quadratic cross function. Important to notice is that the CDA approach displays invariant performance (up to what is most likely due to numerical approximation errors) with respect to the chosen initialisation. This seems not to be the case for the JBSS approach. When theoretical values for the cumulant tensors are used, we observe a slight decrease in performance when the initialisation is taken further away from the global optimum. This may be due to the fact that the multi-linearity of the cumulant tensor is not fully exploited in JBSS, limiting the optimisation to cumulant tensor slices and as such introducing bias in the solution.

When we focus on a limited sample size (10^4 samples), we observe that the CDA approach is completely invariant with respect to the initialisation as parametrised by ε (at least on the studied interval $[-\pi, \pi]$). The JBSS approach now shows an invariant character in the neighbourhood of the solution (up to $|\varepsilon| \approx 3\pi/4$). A striking observation is that the cross-cost optimisation may alter the performance, since its approximate solution can not be corrected by the subsequent gradient approach, as observed in the close vicinity of the optimal solution, Figure C.3(bottom). The poor performance of the gradient approach may be due to the fact that the manifold structure is not respected in its updates, using a re-projected gradient approach.

For limited sample sizes the minimum cost function no longer corresponds to the optimal separator, see Figure C.3(top). Indeed, the variances of the estimators of the fourth order statistics do no longer allow to identify the unmixing matrix with the global minimum of the cost function, although the solution for 10^4 samples has an acceptable -30dB ISI_J performance index.

Figure C.5 clearly shows the influence of the prior on the performances, as well as the performance gain over CCA and JBSS for the case $D_1 = D_2 = I = 3$. The proposed method identifies the unmixing matrix when the variables of interest are available in multiple observation sets that are corrupted by independent non-gaussian noise. Unfortunately, this performance gain is compromised for $I = 2, D_1 = D_2 = 3$, especially with respect to CCA, which might be explained by the fact that a single direction per observation set is non-informative (with respect to the common variable entries of interest). These directions can be linearly separated from the subspace spanned by the two remaining, dependent sources.

As we might expect, it is better to overestimate I, since sources that are not matching up in the different sets do not influence negatively on the end result. This may be explained by zero cross-cumulants, remaining zero under linear combinations within the observation spaces. Note that the model is essentially

under determined and that the estimates of the mixing matrices ($\mathbf{A}^{[k]}$) do not permit to estimate the sources $\mathbf{s}^{[k]}$. Instead, we are limited to the estimates of $\tilde{\mathbf{s}}^{[k]}$.

7.3. Computational Issues

The number of sweeps required for convergence (in distribution) is of the same order of magnitude as $D_1 = D_2$ (Tables C.4 and C.5). This is a small overhead compared to the $2 + \sqrt{D_1} + \sqrt{D_2} = 2(1 + \sqrt{D_1})$ of sweeps generally required by the CoM2 algorithm when ran on each of the sets [6]. This makes the slightly more expensive eigenvalue decomposition of the companion matrix an acceptable investment. The Jacobi algorithm permits a parallel implementation [7, Sec. 8.4.6], allowing for an efficient implementation on multi-core or multi-processor architectures.

7.4. Future Research Directions

Future research will be devoted to a more efficient manifold based optimisation approach [28] that is not based on sweeps over all possible pairs and the extension of the proposed manifold based optimisation approach to the case of non-orthogonal matrices, omitting the prewhitening step as well as to the complex (non-circular) case.

8. Conclusions

In this manuscript we have shown the advantages of Jacobi iterations to solve the problem of Canonical Dependence Analysis. The proposed method has shown excellent performance on the problem of the separation of multivariate Laplace distributed variables. Indeed, the proposed algorithm shows an invariance with respect to its initialisation and outperforms the joint Blind Source Separation on these datasets. However, the main advantage of the proposed method can be found in the identification of the mixing matrices of multiset observations when the variables that are linearly dependent between the sets are corrupted by non-Gaussian noise. We have indeed an empirical proof of the superior performance of the proposed algorithm both with respect to the joint Blind Source Separation and the Canonical Correlation Analysis. The relatively high computational cost when dealing with large dimensions can easily be counterbalanced by the fact that Jacobi algorithms are perfectly well suited for a parallel implementation.

Acknowledgements

The author acknowledges the financial support from the French national research agency (ANR) under the contract ANR-Blanc NT09_511856 (Gaze-EEG), the remarks of the anonymous reviewers and the associate reviewer and the fruitful discussions with Nathalie Guyader and Bertrand Rivet. All have contributed to the current state of the manuscript.

Appendix A. Polynomial coefficients

The contrast function for canonical dependence analysis is maximised by updating the current observations in the first observation space, switching to the second observation space and repeating. We may thus focus on $\mathbf{y}^{[1]}$ without loss of generality. Writing the orthogonal matrix $\mathbf{Q}^{[1]}$ as a series of planar rotations $\mathbf{Q}_{ij}^{[k]}$ and fixing the current index pair to (i, j), we may denote with $\kappa^{\mathbf{x}}_{a,b,c,d} = \kappa^{\mathbf{x}_{[i,j]}}_{a,b,c,d}(\theta^{[1]}_{ij})$ the cumulants of the estimates and by $\kappa^{\mathbf{y}}_{a,b,c,d} = \kappa^{\mathbf{y}_{[i,j]}}_{a,b,c,d}$ the cumulants of the current observations. The planar rotation matrix $\mathbf{Q}_{ij}^{[k]}$ then is the identity matrix, except for the entries (i,i), (i,j), (j,i) and (j,j) which are respectively given by $\cos(\theta^{[1]}_{ij})$, $-\sin(\theta^{[1]}_{ij})$, $\sin(\theta^{[1]}_{ij})$ and $\cos(\theta^{[1]}_{ij})$. The pairwise contrast of order $R = 4$ reads:

$$\begin{aligned}\Psi_{CDA}(\theta^{[1]}_{ij}) &= (\kappa^{\mathbf{x}}_{4,0,0,0})^2 + 4(\kappa^{\mathbf{x}}_{3,0,1,0})^2 + 6(\kappa^{\mathbf{x}}_{2,0,2,0})^2 + 4(\kappa^{\mathbf{x}}_{1,0,3,0})^2 + (\kappa^{\mathbf{x}}_{0,0,4,0})^2 \dots \\ &+ (\kappa^{\mathbf{x}}_{0,4,0,0})^2 + 4(\kappa^{\mathbf{x}}_{0,3,0,1})^2 + 6(\kappa^{\mathbf{x}}_{0,2,0,2})^2 + 4(\kappa^{\mathbf{x}}_{0,1,0,3})^2 + (\kappa^{\mathbf{x}}_{0,0,0,4})^2 \end{aligned} \tag{A.1}$$

The multi-linear relation between the cumulants of the estimates and the cumulants of the current (transformed) observations gives us

$$\begin{cases} \kappa^{\mathbf{x}}_{4,0,0,0} &= \left(\kappa^{\mathbf{y}}_{4,0,0,0} - 4\ t\ \kappa^{\mathbf{y}}_{3,1,0,0} + 6\ t^2\ \kappa^{\mathbf{y}}_{2,2,0,0} - 4\ t^3\ \kappa^{\mathbf{y}}_{1,3,0,0} + t^4\ \kappa^{\mathbf{y}}_{0,4,0,0}\right) / \left(1+t^2\right)^2 \\ \kappa^{\mathbf{x}}_{3,0,1,0} &= \left(\kappa^{\mathbf{y}}_{3,0,1,0} - 3\ t\ \kappa^{\mathbf{y}}_{2,1,1,0} + 3\ t^2\ \kappa^{\mathbf{y}}_{1,2,1,0} - t^3\ \kappa^{\mathbf{y}}_{0,3,1,0}\right) / \left(1+t^2\right)^{3/2} \\ \kappa^{\mathbf{x}}_{2,0,2,0} &= \left(\kappa^{\mathbf{y}}_{2,0,2,0} - 2\ t\ \kappa^{\mathbf{y}}_{1,1,2,0} + t^2\ \kappa^{\mathbf{y}}_{0,2,2,0}\right) / \left(1+t^2\right) \\ \kappa^{\mathbf{x}}_{1,0,3,0} &= \left(\kappa^{\mathbf{y}}_{1,0,3,0} - t\ \kappa^{\mathbf{y}}_{0,1,3,0}\right) / \left(1+t^2\right)^{1/2} \\ \kappa^{\mathbf{x}}_{0,4,0,0} &= \left(\kappa^{\mathbf{y}}_{0,4,0,0} + 4\ t\ \kappa^{\mathbf{y}}_{1,3,0,0} + 6\ t^2\ \kappa^{\mathbf{y}}_{2,2,0,0} + 4\ t^3\ \kappa^{\mathbf{y}}_{3,1,0,0} + t^4\ \kappa^{\mathbf{y}}_{4,0,0,0}\right) / \left(1+t^2\right)^2 \\ \kappa^{\mathbf{x}}_{0,3,0,1} &= \left(\kappa^{\mathbf{y}}_{0,3,0,1} + 3\ t\ \kappa^{\mathbf{y}}_{1,2,0,1} + 3\ t^2\ \kappa^{\mathbf{y}}_{2,1,0,1} + t^3\ \kappa^{\mathbf{y}}_{3,0,0,1}\right) / \left(1+t^2\right)^{3/2} \\ \kappa^{\mathbf{x}}_{0,2,0,2} &= \left(\kappa^{\mathbf{y}}_{0,2,0,2} + 2\ t\ \kappa^{\mathbf{y}}_{1,1,0,2} + t^2\ \kappa^{\mathbf{y}}_{2,0,0,2}\right) / \left(1+t^2\right) \\ \kappa^{\mathbf{x}}_{0,1,0,3} &= \left(\kappa^{\mathbf{y}}_{0,1,0,3} + t\ \kappa^{\mathbf{y}}_{1,0,0,3}\right) / \left(1+t^2\right)^{1/2} \end{cases}$$

where we have used $t = \tan\left(\theta^{[1]}_{ij}\right)$. Furthermore, we have that $(\kappa^{\mathbf{x}}_{0,0,4,0})^2 = (\kappa^{\mathbf{y}}_{0,0,4,0})^2$ and $(\kappa^{\mathbf{x}}_{0,0,0,4})^2 = (\kappa^{\mathbf{y}}_{0,0,0,4})^2$, i.e., they remain invariant under the planar rotation over an angle $\theta^{[1]}_{ij}$. Plugging the above equalities back into the expression for the pairwise contrast function (A.1), we obtain a polynomial in $t = \tan(\theta^{[k]}_{ij})$ of the form

$$\Psi_{CDA}(\theta^{[1]}_{ij}) = \frac{a_8 t^8 + a_7 t^7 + a_6 t^6 + a_5 t^5 + a_4 t^4 + a_3 t^3 + a_2 t^2 + a_1 t + a_0}{(1+t^2)^4} \tag{A.2}$$

where the coefficients a_i are obtained by putting all terms on the common denominator $(1+t^2)^4$. One sees that if no associated data is available for the

current pair of observations (i.e., $I < i < j$ and thus $\mathbf{y}_{[i,j]} = (y_i^{[1]}, y_j^{[1]}, \square_i^{[2]}, \square_j^{[2]})$), the contrast reduces to

$$\Psi_{CDA}(\theta_{ij}^{[1]}) = (\kappa_{4,0,0,0}^{\mathbf{x}})^2 + (\kappa_{0,4,0,0}^{\mathbf{x}})^2 = \Psi_{CoM2}(\theta_{ij}^{[1]})$$

which is exactly the rational polynomial function of the CoM2 contrast [6]. Its minima and maxima can be obtained analytically by zeroing its derivative, since both the denominator and the numerator contain polynomials of degree $2R$ in t having the symmetry $p(t) = t^{2R}p(-t^{-1})$. This means they can be reduced to a polynomial of degree $R(\leq 4)$ in ξ by the change of variable $\xi = t - t^{-1}$. However, if matching signals are available in the associated dataset, the polynomials do no longer exhibit this symmetry and the zeros of the derivative need to be sought through an iterative procedure such as the root finding algorithm based on the eigenvalue decomposition of the companion matrix (see Appendix B).

Appendix B. Companion Matrix

The companion matrix of a monic polynomial $p(x) = a_R(x^R + a_{R-1}x^{R-1} + \ldots + a_2x^2 + a_1x + a_0)$ is given by [7]

$$\mathbf{C}(p) = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 & -a_0 \\ 1 & 0 & 0 & \ldots & 0 & -a_1 \\ 0 & 1 & 0 & \ldots & 0 & -a_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & -a_{R-1} \end{pmatrix}$$

The eigenvalues of this matrix are given by $\{\lambda; \det(\mathbf{C}(p) - \lambda\mathbf{I}_R) = 0\}$. Developing the determinant, we observe that the eigenvalues are indeed the solutions to $p(x) = 0$, as required.

Appendix C. Geodesic Distance on SO(n)

The geodesic distance on a Lie group is given by $d(\mathbf{Q}_1, \mathbf{Q}_2) = \| \log(\mathbf{Q}_1^T\mathbf{Q}_2)\|_F$. The special orthogonal group $SO(n)$ may be represented by the skew symmetric matrices 'Skew(n)' as $SO(n) = e^{\text{Skew}(n)}$. Every skew symmetric matrix has $n(n-1)/2$ degrees of freedom, uniquely determined by its upper (or, lower) triangular elements. As a consequence, a skew symmetric matrix may be represented as $S(\mathbf{r})$, where $\mathbf{r} \in \mathbb{R}^{n(n-1)/2}$ is the vector containing the upper triangular elements. It follows that for an orthogonal matrix $\mathbf{Q}$: $d^2(\mathbf{I}, \mathbf{Q}) = \| \log(\mathbf{Q})\|_F^2 = \| \log(e^{S(\mathbf{r})})\|_F^2 = \|S(\mathbf{r})\|_F^2 = 2\|\mathbf{r}\|_2^2$. Multiplying $\mathbf{r}$ by ε is equivalent to multiplying $S(\mathbf{r})$ by ε and thus results in a distance $\varepsilon\|S(\mathbf{r})\|_F = \varepsilon$, where we have taken $\|S(\mathbf{r})\|_F = 2\|\mathbf{r}\|_2^2 = 1$. If the optimal separation matrix corresponds to $\mathbf{Q}^{[k]} = {\mathbf{A}^{[k]}}^T$, then we have $d(\mathbf{Q}^{[k]}, \mathbf{Q}^{[k]}e^{\varepsilon S(\mathbf{r})})$. For SO(2), we have no choice but to take $\mathbf{r} = r = \pm 1$, which (for $r = 1$):

$$e^{\varepsilon \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)} = \begin{pmatrix} \cos\varepsilon & -\sin\varepsilon \\ \sin\varepsilon & \cos\varepsilon \end{pmatrix} .$$

[1] M. Congedo, E. John, D. De Ridder, L. Prichep, Group independent component analysis of resting-state eeg in large normative samples, International Journal of Psychophysiology 78 (2010) 89–99.

[2] J. Lee, T. Lee, F. Jolesz, S. Yoo, Independent vector analysis (IVA): multivariate approach for fMRI group study, NeuroImage 40 (2008) 86–109.

[3] M. Congedo, R. Phlypo, D.-T. Pham, Approximate joint singular value decomposition of an asymmetric rectangular matrix set, IEEE Transactions on Signal Processing 59 (2011) 415–424.

[4] X.-L. Li, T. Adalı, M. Anderson, Joint blind source separation by generalised joint diagonalization of cumulant matrices, Signal Processing 91 (2011) 2314–2322.

[5] X.-L. Li, M. Anderson, T. Adalı, Second and higher-order correlation analysis of multiple multidimensional variables by joint diagonalization, in: V. Vigneron et al. (Ed.), LVA/ICA 2010, volume 6365 of *LNCS*, pp. 197–204.

[6] P. Comon, Independent component analysis, a new concept?, Signal Processing 36 (1994) 287–314.

[7] G. H. Golub, C. F. Van Loan, Matrix Computations, The Johns Hopkins University Press, 3rd edition, 1996.

[8] H. Hotelling, The most predictable criterion, Journal of Educational Psychology 26 (1935) 139–142.

[9] H. Hotelling, Relation between two sets of variates, Biometrika 28 (1936) 321–377.

[10] T. M. Cover, J. A. Thomas, Elements of Information Theory, Wiley Interscience, 2nd edition, 2006.

[11] J. D. Carroll, Generalization of canonical correlation analysis to three or more sets of variables, in: Proceedings of the 76th Annual Convention APA, pp. 227–228.

[12] Y. Li, T. Adalı, W. Wang, V. Calhoun, Joint blind source separation by multiset canonical correlation analysis, IEEE Trans on Signal Processing 57 (2009) 3918–3929.

[13] J. R. Kettenring, Canonical analysis of several sets of variables, Biometrika 58 (1971) pp. 433–451.

[14] R. Phlypo, M. Congedo, An extension of the canonical correlation analysis to the case of multiple observations of two groups of variables, in: Proceedings of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Buenos Aires, Argentina.

[15] T. Kim, H. T. Attias, S. Lee, T. Lee, Blind source separation exploiting higher-order frequency dependencies, IEEE Trans on Speech and Audio Processing 15 (2007) 70–79.

[16] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, R. C. Goodlin, Adaptive noise cancelling: Principles and applications, Proceedings of the IEEE 63 (1975) 1692–1716.

[17] R. Phlypo, V. Zarzoso, I. Lemahieu, Source extraction by maximising the variance in the conditional distribution tails, IEEE Trans on Signal Processing 58 (2010) 305–316.

[18] W. Lu, J. Rajapakse, Constrained independent component analysis, in: Advances in Neural Information Processing Systems, volume 13, MIT Press, 2000, pp. 570 – 576.

[19] W. Lu, J. C. Rajapakse, Approach and applications of constrained ICA, IEEE Trans on Neural Networks 16 (2005) 203–212.

[20] A. Hyvärinen, E. Oja, One-unit learning rules for independent component analysis, in: M. Mozer, M. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems, volume 9, Cambridge: MIT Press, 1997, pp. 480 – 486.

[21] M. Sato, Y. Kimura, S. Chida, T. Ito, N. Katayama, K. Okamura, M. Nakao, A novel extraction method of fetal electrocardiogram from the composite abdominal signal, IEEE Trans on Biom Eng 54 (2007) 49–58.

[22] R. Phlypo, V. Zarzoso, I. Lemahieu, Atrial activity estimation from atrial fibrillation ECGs by blind source extraction based on a conditional maximum likelihood approach, Medical & Biological Engineering & Computing 48 (2010) 483–488.

[23] A. Stuart, K. Ord, Distribution Theory, volume 1 of *Kendall's advanced theory of statistics*, Hodder Arnold, 6th edition, 2006.

[24] P. Mc Cullagh, Tensor Methods in Statistics, Chapman and Hall, 1987.

[25] E. Moreau, O. Macchi, A one stage self-adaptive algorithm for source separation, in: IEEE Int Conf on Acoustics, Speech and Signal Proc, volume 3, pp. 49–52.

[26] S.-I. Amari, A. Cichocki, H. H. Yang, A new learning algorithm for blind signal separation, in: Advances in Neural Information Processing Systems, volume 8, MIT Press, Cambridge MA, 1996, pp. 757–763.

[27] T. Itahashi, K. Matsuoka, Stability of independent vector analysis, Signal Processing (2011). Doi:10.1016/j.sigpro.2011.11.008.

[28] P.-A. Absil, R. Mahony, R. Sepulchre, Optimization Algorithms on Matrix Manifolds, Princeton University Press, Princeton, NJ, 2008.

List of Figures

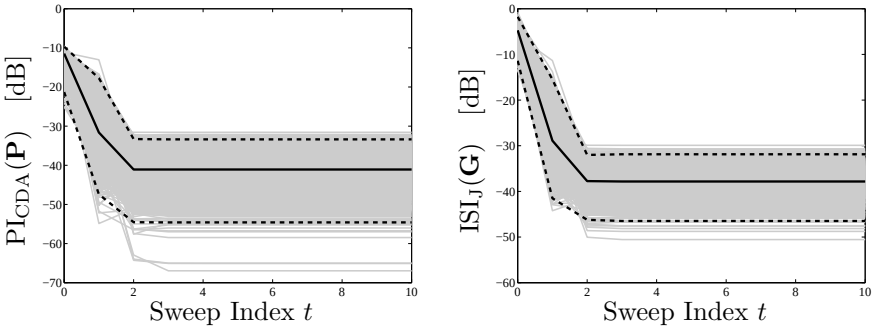

Figure C.1: The evolution of the performance indices $\mathrm{PI}_{\mathrm{CDA}}$ of Equation (8) and $\mathrm{ISI}_{\mathrm{J}}$ of Equation (9) as a function of the sweep index t. The observations are based on 10^4 samples of two linear mixtures of Multivariate Laplace distributed variables (see subsection 6.2 for details). The results of each of the 1000 Monte Carlo realisations are represented by a light gray line and the overlay contains the median (black, continuous), the 1st and the 99th percentile (black, dashed).

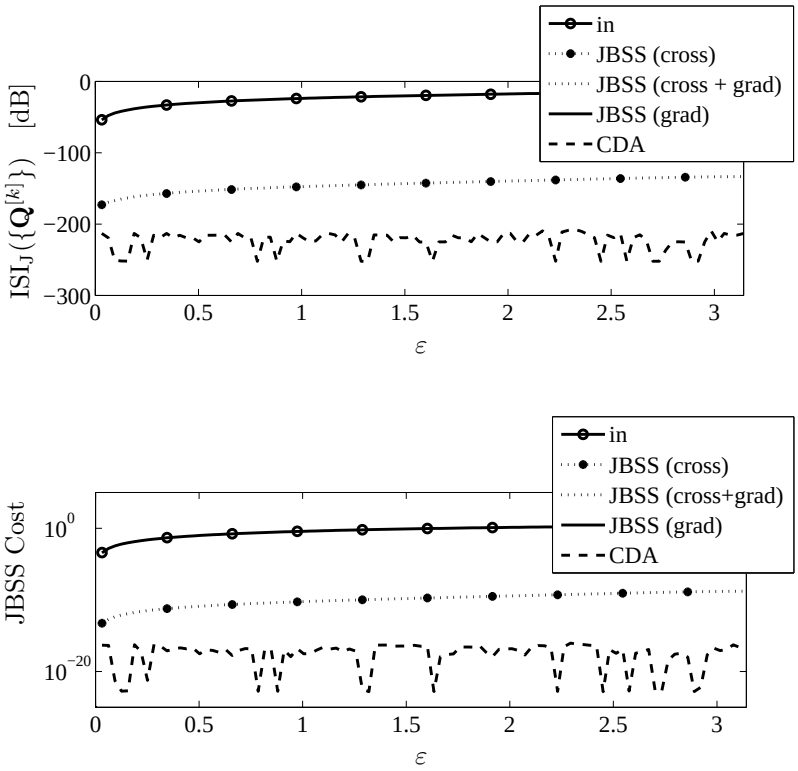

Figure C.2: (top) Performance measure and (bottom) JBSS cost function values at convergence when the algorithms take the theoretical cumulant values as an argument. Perturbation of the optimal solution is obtained by displacing the initialisation of the algorithms over a distance ε on SO(3). Results are taken as the mean over 100 Monte Carlo runs.

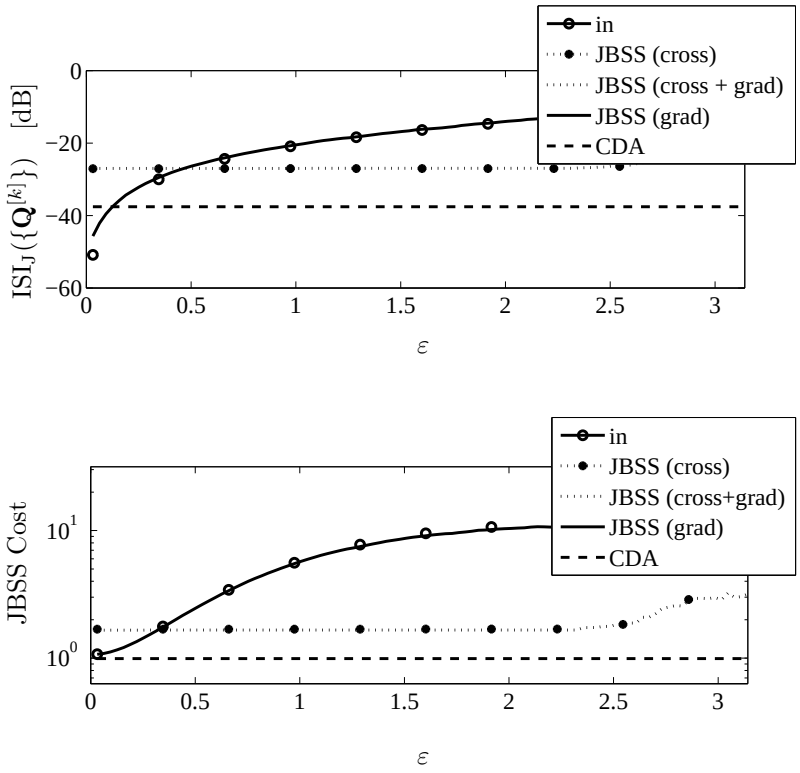

Figure C.3: (top) Performance measure and (bottom) JBSS cost function values for a sample size of 10^4. Perturbation of the optimal solution is obtained by displacing the initialisation of the algorithms over a distance ε on SO(3). Results are taken as the mean over 100 Monte Carlo runs.

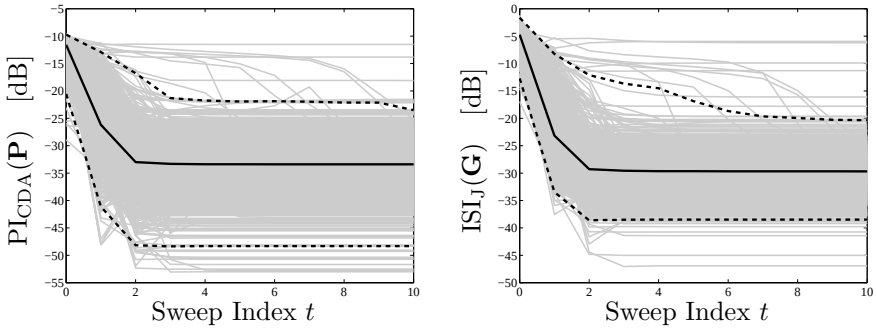

Figure C.4: The evolution of the performance indices $\mathrm{PI}_{\mathrm{CDA}}$ of Equation (8) and $\mathrm{ISI}_{\mathrm{J}}$ of Equation (9) as a function of the sweep index t. There are three observations for three nuisance sources per observation set ($\nu = 1$) with three shared variables between the sets, see subsection 6.3 for the generative model. The results of each of the 1000 Monte Carlo realisations are represented by a light gray line and the overlay contains the median (black, continuous), the 1st and the 99th percentile (black, dashed).

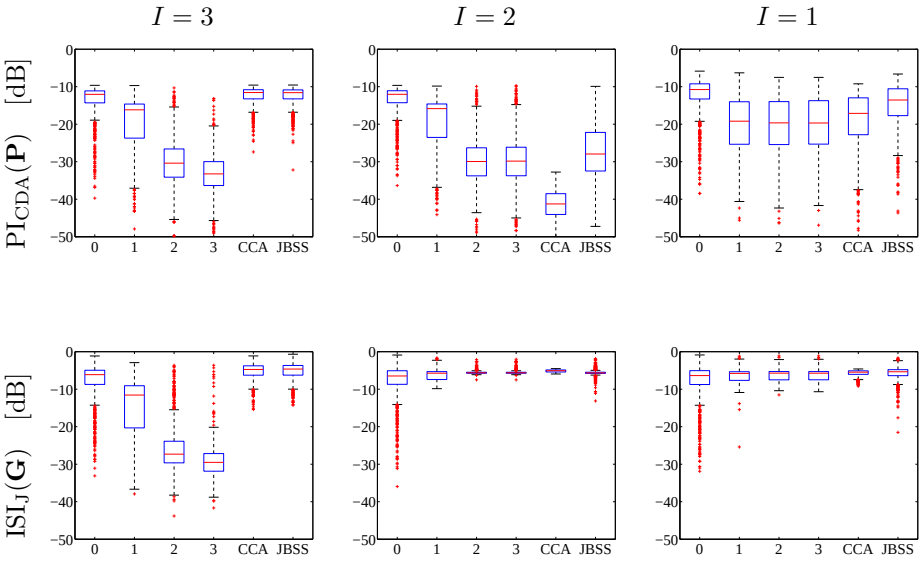

Figure C.5: The performance measured through $\mathrm{PI}_{\mathrm{CDA}}(\mathbf{P})$ of Eq. (8) and $\mathrm{ISI}_{\mathrm{J}}$ of Eq. (9) as a function of the algorithm ($0, 1, 2, 3$: CDA with estimated number of common sources $\hat{I} = 0, 1, 2, 3$, JBSS [4] and CCA [9]) and the actual number of matched sources I. Each set is composed of I sources common to both observation sets, $(3 - I)$ sources proper to each of the observation sets and 3 nuisance sources, according to model (10) with $\nu = 1$. The results are given in box-whiskers plots, containing the 25th and 75th percentile as the box extremities and the median as the horizontal line within the box. Outliers (crosses) are those samples that are outside the interval $[\mathrm{pct}_{25} - \frac{3}{2}(\mathrm{pct}_{75} - \mathrm{pct}_{25}),\ \mathrm{pct}_{75} + \frac{3}{2}(\mathrm{pct}_{75} - \mathrm{pct}_{25})]$, where pct_c is the c% percentile. Data has been collected over 1000 Monte Carlo runs.

List of Tables

Table C.1: Pseudocode to calculate the optimal rotation matrices $\mathbf{Q}^{[k]}$ such that $\mathbf{Q} = \text{blkdiag}(\mathbf{Q}^{[1]},\ \mathbf{Q}^{[2]},\ \ldots,\ \mathbf{Q}^{[K]}) = \widehat{\mathbf{MA}^{-1}}$

initialize $\mathbf{Q}^{[k]} = \mathbf{I}_{M_k}$
while no convergence **do**
  for $k = 1 \rightarrow K$ **do**
    for $i = 1 \rightarrow \max_k(D_k) - 1$ **do**
      for $j = i + 1 \rightarrow \max_k(D_k)$ **do**
        calculate $\widehat{\theta_{ij}^{[k]}} = \arg\max_{\theta_{ij}^{[k]}} \Psi_{CDA}(\theta_{i,j}^{[k]})$
        $= \sum_{r_1+r_2=R} \left(\kappa^{\mathbf{x}}_{r_1,0,r_2,0}(\theta_{ij}^{[k]})\right)^2 + \left(\kappa^{\mathbf{x}}_{0,r_1,0,r_2}(\theta_{ij}^{[k]})\right)^2$
        (see Appendix A, Eq. A.2)
        update the observations as $\mathbf{x}_{[i,j]} \leftarrow \mathbf{x}_{[i,j]}(\widehat{\theta_{ij}^{[k]}})$
        update $\mathbf{Q}^{[k]} \leftarrow \mathbf{Q}_{ij}^{[k]}(\widehat{\theta_{ij}^{[k]}})\ \mathbf{Q}^{[k]}$
      end for
    end for
  end for
end while

Table C.2: The CDA algorithm

Require: Datasets $\mathbf{X}^{[k]}$ of equal samples size N

return Canonical Dependence Components $\mathbf{Y}^{[k]}$

Unmixing matrices $\mathbf{W}^{[k]}$, such that $\mathbf{X}^{[k]} = \mathbf{A}^{[k]}\mathbf{Y}^{[k]}$

for $k = 1 \rightarrow K$ **do**

{make the observations zero-mean and pre-whiten on a dataset level ($\mathbb{1}_N$ is a column vector of N ones)}

$\mathbf{X}^{[k]} \leftarrow \mathbf{X}^{[k]} - \boldsymbol{\mu}^{[k]}\mathbb{1}_N^T$

Compute SVD $\mathbf{X}^{[k]} = \mathbf{U}\boldsymbol{\Sigma}^{[k]}\mathbf{V}^T$

$\mathbf{X}^{[k]} \leftarrow \sqrt{N}\left(\boldsymbol{\Sigma}^{[k]}\right)^{-1}\left(\mathbf{U}^{[k]}\right)^T\mathbf{X}^{[k]}$

end for

Estimate the rotation matrices $\mathbf{Q}^{[k]}$ using the algorithm in Table C.1

$\mathbf{Y}^{[k]} = \mathbf{X}^{[k]}$ {$\mathbf{Y}$ is equal to $\mathbf{X}$ at the last iteration}

$\mathbf{A}^{[k]} = \mathbf{U}^{[k]}\boldsymbol{\Sigma}^{[k]}\left(\mathbf{Q}^{[k]}\right)^T$

Table C.3: A summary of the experiments and the results that can be found in this manuscript. $\mathcal{L}$ap and $\mathcal{U}$ni denote respectively Laplacian and Uniformly distributed variables. Details about the experiments can be found in the referred sections. D_1 and D_2 refer to the dimensions of the datasets, I to the number of sources they have in common.

section	results	D_1	D_2	I	noise	distribution	sample size	objective
6.2.1	Fig. C.1	3	3	3	no	$\mathcal{L}$ap	10^4	convergence
6.2.1	Table C.4	I	I	$2, 3, 5, 10$	no	$\mathcal{L}$ap	10^4	convergence
6.2.2	Fig. C.2	3	3	3	no	$\mathcal{L}$ap	∞	performance
6.2.3	Fig. C.3	3	3	3	no	$\mathcal{L}$ap	10^4	performance
6.3.1	Fig. C.4	3	3	3	yes	$\mathcal{L}$ap/$\mathcal{U}$ni	10^4	convergence
6.3.1	Table C.5	I	I	$2, 3, 5, 10$	yes	$\mathcal{L}$ap/$\mathcal{U}$ni	10^4	convergence
6.3.2	Fig. C.5	3	3	$1, 2, 3$	yes	$\mathcal{L}$ap/$\mathcal{U}$ni	10^4	performance

Table C.4: The p-values issued from a Kolmogorov-Smirnov test on the similarity between the distributions of the performance indices over two consecutive sweep numbers $t-1$ and t. Data are drawn from multivariate Laplace distributions as described in subsection 6.2. The table is truncated at the seventh iteration, since convergence has been reached for all dimensions.

I \ t	2	3	4	5	6	7
2	0	$2.63\ 10^{-3}$	**1.00**	**1.00**	**1.00**	**1.00**
3	0	0.00	**1.00**	**1.00**	**1.00**	**1.00**
5	0	0	$6.55\ 10^{-9}$	**1.00**	**1.00**	**1.00**
10	0	0	0.00	**1.00**	**1.00**	**1.00**

Table C.5: The p-values issued from a Kolmogorov-Smirnov test on the similarity between the distributions of the performance indices over two consecutive sweep numbers $t-1$ and t. Data are drawn from the non-Gaussian noise model defined in subsection 6.3. The table is truncated at the eighth iteration, since convergence has been reached for all dimensions.

I \ t	2	3	4	5	6	7	8
2	0.00	$1.65\ 10^{-7}$	$9.93\ 10^{-1}$	**1.00**	**1.00**	**1.00**	**1.00**
3	0	$3.35\ 10^{-96}$	$1.17\ 10^{-1}$	**1.00**	**1.00**	**1.00**	**1.00**
5	0	0.00	$4.11\ 10^{-16}$	$3.07\ 10^{-1}$	**1.00**	**1.00**	**1.00**
10	0	0.00	0.00	$4.39\ 10^{-5}$	$5.65\ 10^{-1}$	$9.96\ 10^{-1}$	**1.00**